International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Survey Paper on Slicing Concept Used for Privacy Preserving

Vishal Chavan[1], Ravindra Pal[2], Jitendra Pal[3]

[1, 2, 3]VIVA Institute of Technology, Shirgaon, Virar (East), Maharashtra, India

Abstract: ***Many techniques has been design for privacy preserving and micro data publishing such as generalization, Bucketization a. But there is problem in generalization i.e it is not suitable for lagre amount of data and there is a problem in Bucketization is that, it does not maintain a membership disclosure and it is also not suitable where there is no clear separation between quasi identifier and sensitive attribute. And both this problem is being solved in slicing Slicing uses a combination of both generalization and bucketization in order to preserve the privacy of data.***

Keywords: Data Anonymization, Data publishing, Generalization, Bucketization**,** K-anonymity, T-Closeness, Slicing

1. Introduction

In today's world the gathering of digital information by various agencies has given a huge amount of opportunities for knowledge based decision making. Privacy has become an important aspect, during publishing the data to the outside world. There are various organization which are required to published the data. For instance, authorized healing centers in California are required to submit particular statistic information each patient released from their office. The data which is going to be published contains a detailed person specific data which contains sensitive attribute that can give specific information about individual which can directly violates a person privacy. Microdata is the type of data which is going to be published in the form of records. These data is classified into three categories namely, 1) explicit identifier: it provides clear identification about an individual such as name address. 2) quasi identifier: the attribute whose values are combined together can give the identification of an individuals.3)Sensitive attribute : the attributes whose values can give the identification of a person.

In the earlier years we are contemplating Privacy-preserving and publishing of micro data. Micro data contains records of every data around an individual, for example, a man, a division, or an organization. Various data anonymization techniques are present. Data anonymization is the encrypting of data for privacy preservation. There are three sorts of disclosure exist. Those are attribute disclosure, identity disclosure and membership disclosure. Identity disclosure means when information about some individual is leaked. It occurs a particular record associated with some individual is released. Attribute disclosure occurs when information some individual is exposed. Due to identity disclosure attribute disclosure happens. When the information about certain group or organization is leaked it is known as membership disclosure, it is a specific type of attribute disclosure.

Table 1: Original Table

Age	Work class	Occupation	Sex	Native country
30	State government	Exec- manager	M	United- state
38	private	Handler- cleaner	F	Jamaica
35	private	Handler- cleaner	M	United- state
28	private	Exec- manager	M	Cuba
50	Local government	sales	F	Jamaica
49	private	sales	F	United- state
53	State government	Exec- manager	M	Cuba
51	State government	sales	M	Cuba

2. Literature Survey

2.1 K-anonymity

The database where attributes are suppressed or generalized until each row is indistinguishable with in any event k-1 different lines that database is said to be K-anonymous. K-anonymity show for numerous sensitive attributes said that there are three sorts of information disclosure. A comparability class is said to be l-diverse in the event that it has utilized most extreme number of l "well-represented" qualities in understanding of sensitive attribute.

2.2 T-Closeness

In these the privacy measurement is given by the information gained to the intruder through the revealed data. Information gain means the information or the knowledge learned by the intruder. The intruder has some belief before observing the data after receiving the data the intruder has some new knowledge about the data and its belief is changed from prior to posterior belief. Information gain is given by calculating the difference between the prior belief and posterior belief.If the distance between the scattering of a sensitive attribute in this class and the scattering of the attribute in the whole table is no more than a threshold t then an equivalence class is said to have t-closeness. If all equivalence classes have t- closeness A table is said to have t-closeness.

2.3 Generalization

Generalization is an anonymizaion it can be implemented using k-anonymity.k-anonymity has been described in the

upper section. In generalization the attributes are crubed until each row is identical. It changes the quasi identifier in every bucket to a less specific but into semantically constant values. In these the rows are effected. It is better than th previous technique. It has a drawback that it does not provide better data utilization.

Table 2: Generalization table

Age	Work class	Occupation	Sex	Native country
[28-38]	State government	Exec-manager	*	United-state
[28-38]	private	Handler-cleaner	*	Jamaica
[28-38]	private	Handler-cleaner	*	United-state
[28-38]	private	Exec-manager	*	Cuba
[49-53]	Local government	sales	*	Jamaica
[49-53]	private	sales	*	United-state
[49-53]	State government	Exec-manager	*	Cuba
[49-53]	State government	sales	*	Cuba

2.4 Bucketization

The next anonymization technique brought was bucketization. It was brought in order to overcome the drawbacks of generalization. It gives better data utilization. In bucketization the data is divided into sensitive attribute amd quasi identifier. It is achieved by randomly permuting the SAs values from QI.In bucketization one column has only SAs while the other column keeps QI attributes.

Table 3: Bucketization table

Age	Work class	Occupation	Sex	Native country
30 38 35 28	State government Private Private Private	Exec-manager Handler-cleaner Handler-cleaner Exec-manager	M F M M	Jamaica United-state United-state Cuba
50 49 53 51	Local government Private State government State government	Sales Sales Exec-manager sales	F F M M	Jamaica Cuba Cuba United-states

2.5 Slicing

A new accentuate technique is developed for data publishing with data privacy and data utility is called as slicing. In slicing first, vertically partition attribute in the table into columns. Each column consists of a subset of attribute. So that highly correlated attributes value in same column and preserve correlation among those attribute this is good for data utility and breaking association of uncorrelated attribute it is also good for data privacy because the association between values of uncorrelated is much less frequent and thus more identifiable and noticeable to adversary. Slicing also, horizontally partition tuples in the table into buckets. Each bucket consists of a subset of tuples. Then after within each bucket, values in each column are randomly permuted and then break correlation between values of different columns. So that the correlation between the values of two columns within one bucket is hidden from adversary Multiset based generalization is equivalent to a trivial slicing approach where each column contains exactly one attribute.

3. Review of generalization and bucketization

3.1 Problem Statement

Privacy preserving data publishing is an issue now days. While data get published to any agencies, there is risk of information disclosure. While reducing information disclosure risk there is loss of data utility. Generalization and bucketization may fail to achieve data privacy and utility because during attribute partitioning sensitive attribute is grouped into single column. Since there is considerable amount of information lost for a high dimensional data in generalization and bucketization does not maintain membership disclosure.

3.2 Proposed Technique

The proposed technique is slicing in which attributes separated by using bot generalization and bucketization. Slicing partitions attribute both horizontally and vertically. In vertical partitioning more correlated attributed are taken into one group and uncorrelated attributed are grouped separately. In horizontal partitioning tuple are grouped to form buckets, after grouping tuples values of column are randomly permuted. slicing works in two main steps.

1. Attribute partitioning
2. Tuple partitioning

3.3 Attribute partitioning

In attribute partitioning, correlation of the attribute are measured to form there group. To measure the correlation mean square contingency coefficient is used. Mean square coefficient is achieved by following formula:

3.4 Algorithm Partitioning Around Medoid (PAM)

Randomly select k of the n data points as the medoid

1) Associate each data point to the closest medoid.
2) For each medoid m 1. For each non-medoid data point o . swap m and o and compute the total cost of the configuration
3) Select the configuration with the lowest cost
4) Repeat steps 2 to 4 until there is no change in the medoid.

3.5 Tuple partitioning

The algorithm maintains two data structures: (1) a queue of buckets Q and (2) a set of sliced buckets SB. Initially, Q

contains only one bucket which includes all tuples and SB is empty (line 1). In each iteration (line 2 to line 7), the algorithm removes a bucket from Q and splits the bucket into two buckets (the split criteria is described in Mondrian [17]). If the sliced table after the split satisfies ℓ-diversity (line 5), then the algorithm puts the two buckets at the end of the queue Q (for more splits, line 6). Otherwise, we cannot split the bucket anymore and the algorithm puts the bucket into SB (line 7). When Q becomes empty, we have computed the sliced table. The set of sliced buckets is SB (line 8).

4. Expected Result

We evaluate the effectiveness of slicing in preserving data utility and protecting against attribute disclosure, identity disclosure and membership disclosure as compared to generalization and bucketization.

Data set

Some preprocessing steps must be applied on the anonymized data before it can be used for workload tasks. First, the anonymized table computed through generalization contains generalized values, which need to be transformed to some form that can be understood by the classification algorithm. Second, the anonymized table computed by bucketization contains multiple columns, the linking between which is broken. We need to process such data before workload experiments can run on the data.

Handling generalized values

In this step, we map the generalized values (set/interval) to data points. the Mondrian algorithm assumes a total order on the domain values of each attribute and each generalized value is a sub-sequence of the total-ordered domain values. There are several approaches to handle generalized values. The first approach is to replace a generalized value with the mean value of the generalized set. For example, the class 9th, 10th and 11th replaced by 10th. The second approach is to replace a generalized value by its lower bound and upper bound. In this approach, each attribute is replaced by two attributes, doubling the total number of attributes. For example, the Education attribute is replaced by two attributes Lower-Education and Upper Education; for the generalized Education level {9th, 10th,11th}, the Lower Education value would be 9th and the Upper-Education value would be 11th. We use the second approach in our experiments. Handling bucketized/sliced data. In both bucketization and slicing, attributes are partitioned into two or more columns. For a bucket that contains k tuples and c columns, we generate k tuples as follows. We first randomly permutated the values in each column. Then, we generate the ith ($1 \leq i \leq k$) tuple by linking the i-th value in each column. We apply this procedure to all buckets and generate all of the tuples from the bucketized/sliced table. This procedure generates the linking between the two columns in a random fashion. Table 1 contains the record of original micro data table in which Name is explicite identifier which removed in first step. Age, gender, and zip code are quasi identifier and remaining two diseases and occupation are the sensitive attribute. Table 2 is overlapped slicing table in which explicit identifier Name is removed from table and quasi identifier are grouped together with one sensitive attribute and another group of sensitive attribute. The value of sensitive attribute is randomly permuted to achieve more privacy. Sensitive attributes are partitioned with both attribute therefore more attribute correlation is achieved and utility of data is increased.

5. Conclusion

Anonymization technique is powerful method for privacy preserving of published data. This paper presents a new anonymization method that is slicing with new privacy model i.e. by combining both bucketization and generalization. This method overcomes the limitations of slicing and preserves better utility while protecting against privacy threats. slicing is used to prevent attribute disclosures. The general methodology of this work is before data anonymization one can analyze the data characteristics in data anonymization. The basic idea is one can easily design better anonymization techniques when we know the data perfectly. Finally, we have some advantages slicing comparing with generalization and bucketization. Slicing is a promising technique for handling high dimensional data. By increasing the correlation among data privacy is preserved

References

[1] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, Slicing: A New Approach to Privacy Preserving Data Publishing, IEEE 2012 Transactions on Knowledge and Data Engineering, volume:24,Issue:3
[2] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and _-Diversity, Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference
[3] C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, pages 901–909, 2005.
[4] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In KDD, pages 767–775, 2008.
[5] DR.K Swathi and S. Ramanthan, R.Vijyanathan Ordered Bucketization Encryption for the Privacy Secure Data IEEE, p 2, 2006.
[6] Sampler Hang J. Kim & Steven N. MacEachernHang The Generalized Multiset IJRE, p 5, 2009.
[7] I. Dinur and K. Nissim, Revealing Information while Preserving Privacy Proc. CM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.
[8] G.Ghinita, Y.Tao, and P. Kalnis, “On the Anonymization of Sparse High Dimensional Data,” Proc. IEEE 24th Int‟l Conf. Data Eng. (ICDE), pp. 715-724, 2008.
[9] B.C.M. Fung, K. Wang, and P.S. Yu, “Top-Down Specialization for Information and Privacy Preservation,” Proc. Int‟l Conf. Data Eng.(ICDE), pp. 205- 216, 2005.

Author Profile

Vishal Chavan is pursuing B.E. degree from Mumbai University.

Ravindra Pal is pursuing B.E. degree from Mumbai University.

Jitendra Pal is pursuing B.E. degree from Mumbai University.